Nonparametric Regression with Discrete Covariate and Missing Values

BY Song Xi Chen[a,b] and Cheng Yong Tang[c] [1]

[a] Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100871, China
[b] Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA
and
[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117543.

SUMMARY

We consider nonparametric regression with a mixture of continuous and discrete explanatory variables where realizations of the response variable may be missing. An imputation based nonparametric regression estimator is proposed. We show that the proposed approach leads to a leading order variance benefit, whereas smoothing the categorical variables gives a second order variance improvement. We also demonstrate the applications of the proposed approach through numerical simulations and two practical examples.

Keywords: Nonparametric Regression; Discrete kernel smoothing; Imputation; Missing Values; Variance Reduction.

1 Introduction

Nonparametric regression with continuous covariate has been extensively explored in the last two decades; see Härdle (1990) and Fan and Gijbels (1996) for comprehensive overviews. In practice, discrete covariate from qualitative observations often arises from various scenarios including many social and economic studies. Discrete kernel smoothing for nonparametric distribution estimation was originally proposed in Aitchison and Aitken (1976). Its theoretical properties are evaluated in Hall (1981) and Hall and Wand (1988). Recently, there are increasing interests of nonparametric inference with mixed continuous and discrete covariates. These include Li and Racine (2003) for distribution function estimation, Hall, Racine, and Li (2004) for conditional density estimation and variable selection; see Li and Racine (2007) for overviews in econometrics.

Missing values are often encountered in practice. Adequately incorporating data missing mechanism is important for valid and efficient inferences (Little and Rubin, 2002). For parametric inference, the impact of missing data is well documented. For overviews and recent

[1] Corresponding Author: Cheng Yong Tang. Tel: 65-65168949. Fax: 65-68723919. Email: statc@nus.edu.sg. We thank the Editor, AE and the referee for their insightful comments and constructive suggestions. We thank US Census Bureau for support and allowing access to the ACE data. The authors acknowledge support by National Science Foundation grants SES-0518904 and DMS-0604563. Tang also acknowledges support from National University of Singapore Academic research grants.

developments, see Little and Rubin (2002), Robins et al. (1995), Wang and Chen (2009) and Qin, Zhang, and Leung (2009) and references therein. For nonparametric inference with missing values, existing works are mainly confined to situation with only continuous random variables. These include Titterington and Mill (1983) on various imputation schemes for density estimation, Cheng (1994), Chu and Cheng (1995) and Gonzalez-Manteiga and Perez-Gonzalez (2004) on the conditional mean imputation in estimating regression function and overall mean of the response variable; see also Müller (2009) and Efromovich (2011) for recent development.

In this paper, we consider nonparametric regression with both continuous and discrete covariates, say X, where the response variable Y may be missing. We start from studying a nonparametric regression that ignores the portion of data with missing values, which we call complete-cases based estimator. Here "complete" means non-missing values which is a convention in the missing value literature. Although this estimator is still consistent, its efficiency is affected by an inflation of the variance due to its ignoring the incomplete portion of data with missing values. To improve the estimation efficiency, we propose an imputation based nonparametric regression estimator that accommodates both continuous and discrete covariates. Although the variances of both estimators have the same order of magnitude, the imputation based estimator has a smaller variance in the leading order than the complete-cases based estimator, which relates to a feature in approaches of Cheng and Peng (2006) and Cheng et al. (2007). We also conduct a second order asymptotic analysis, which reveals that smoothing the discrete covariates leads to an desired variance reduction in the second order. A comprehensive application of the proposed approach has been implemented in the case study reported in Chen, Tang, and Mule (2010) for the US Census dual system accuracy and coverage evaluation where the main concern is on the population size estimation. The focus of this paper, however, is on properties of nonparametric regression with mixed covariates in the presence of missing values.

The paper is organized as follows. Section 2 outlines the nonparametric kernel regression estimators with mixed covariates. The issue of missing values and their imputation are considered in Section 3. Section 4 reports a theoretical analysis on the imputation based nonparametric regression. Construction of confidence band and data adaptive smoothing are also considered. Simulation results are reported in Section 5 followed by data analyses in Section 6. All the technical proofs are deferred to the Appendix.

2 Nonparametric Regression with mixed covariates

Let $\{(X_i, Y_i)\}_{i=1}^n$ be independent and identically distributed random vectors such that $X_i = (X_{i1}, \cdots, X_{id})$ is a d-dimensional explanatory variable and Y_i is a univariate response. Without loss of generality, we write $X_i = (X_i^c, X_i^u)$ where X_i^c consists of d_c continuous covariates while X_i^u consists of d_u unordered categorical ones, where $d_c + d_u = d$. Our target is to estimate the nonparametric regression function

$$m(x) = E(Y_i|X_i = x),$$

with heteroscedastic conditional variance $\sigma^2(x) = var(Y_i|X_i = x)$. Here $x = (x^c, x^u)$ inherits the same partition of X_i.

To smooth continuous covariates, we apply a d_c-dimensional kernel K which is a radially symmetric probability density function in R^{d_c}. We define $K_h(u) = h^{-d_c}K(u/h)$. Here h is a bandwidth that controls the amount of smoothness. For simplicity in presentation, we consider the product kernel in our studies, i.e. $K(u) = \prod_{i=1}^{d_c} K_1(u_i)$ for $u = (u_1, \cdots, u_{d_c})$, where $K_1(\cdot)$ is some symmetric univariate density function. The product kernel can be generalized to a multivariate kernel by the formulation in Scott (1992).

Smoothing categorical variables is designed to efficiently utilize data information from some neighborhood that may share similar characteristics with the target; see Aitchison and Aitken (1976), Hall (1981) and and Hall et al. (2004). For smoothing unordered categorical covariates, we use the discrete kernel originally proposed by Aitchison and Aitken (1976). Other formulation of discrete kernels is available in Li and Racine (2007). Suppose X_{ij}^u, the j-th component of X_i^u, takes c_j discrete values in $\{0, 1, ..., c_j - 1\}$. The bandwidth for smoothing X_{ij}^u is λ_j and the kernel weight at a x_j^u is

$$\lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \neq x_j^u)$$

where $I(\cdot)$ is the indicator function. As shown in the Appendix, this way of discrete kernel smoothing has advantage in quantifying the second order reduction in variance. The bandwidth λ_j takes values within $[c_j^{-1}, 1]$. Assigning $\lambda_j = c_j^{-1}$ leads to a uniform weight irrespective to the difference between X_{ij}^u and x_j^u, whereas $\lambda_j = 1$ gives a kernel weight of 1 if $X_{ij}^u = x_j^u$ and zero otherwise which coincides with the standard frequency weight. The other λ_j values between c_j^{-1} and 1 offer a range of choices for information combining for efficiency improvement. The kernel used to smooth the entire categorical component $X_i^u = (X_{i1}^u, \cdots, X_{id_u}^u)$ at $x^u = (x_1^u, \cdots, x_{d_u}^u)$ is

$$L(x^u, X_i^u; \vec{\lambda}) = \prod_{j=1}^{d_u} \{\lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \neq x_j^u)\}, \tag{2.1}$$

where $\vec{\lambda} = (\lambda_1, \cdots, \lambda_{d_u})$ is a bandwidth vector. Finally, the overall kernel weight drawn from $X_i = (X_i^c, X_i^u)$ for local estimation at $x = (x^c, x^u)$ is

$$K_h(x^c - X_i^c) L(x^u, X_i^u; \vec{\lambda}).$$

The kernel estimator of $m(x)$ in the absence of missing values is then given by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda}) Y_i}{\sum_{i=1}^{n} K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda})}. \tag{2.2}$$

This is a Nadaraya-Watson type estimator carrying out weighted average of the responses Y_i locally, which allows estimating $m(x)$ without assuming a parametric model.

3 Incorporating Missing Values

Now we consider the situation when the response Y_i can be missing. Let δ_i be the missing indicator of Y_i such that $\delta_i = 0(1)$ for missing (observed) Y_i. We concentrate in this study that the Y_i is missing at random (MAR) (Rubin, 1976). MAR is an important notion in missing data analysis. It means that conditioning on the covariate X_i, the missing mechanism of Y_i is independent of Y_i. In other words, δ_i and Y_i are conditionally independent given X_i, i.e.

$$P(\delta_i = 1 | Y_i, X_i) = P(\delta_i = 1 | X_i) =: w(X_i). \tag{3.1}$$

Here, $w(\cdot)$ is called the missing propensity function of Y_i. MAR is weaker and more general than the so called missing completely at random which assumes that the propensity $w(x)$ is a constant function. The statistical inference in case of not missing at random often relies on some parametric model assumption, see for instance the study in (Qin, Leung, and Shao, 2002). In this paper, we do not rely on any parametric model assumption on the missing propensity function. Instead, we discuss practical remedy to make the MAR assumption reasonable based on available information from data.

Here we extend the MAR assumption (3.1) to the following form to reflect the reality of applications by assuming that in addition to X_i, some extra covariate Z_i contributes to the missingness of Y_i. Specifically, we assume that Y_i is missing at random given (X_i, Z_i), i.e.

$$P(\delta_i = 1 \mid Y_i, X_i, Z_i) = P(\delta_i = 1 \mid X_i, Z_i) =: w(X_i, Z_i). \tag{3.2}$$

The rationale for (3.2) is based on practical considerations to make (3.1) more accommodative. As a concrete example in the US Census dual system surveys, in addition to a set of covariates X_i which consists of racial origin, age, sex, housing tenure and region, an operational variable Z_i called match coding group collected during the non-response follow-up process is also predictive to the data missingness as shown in Belin et al. (1993). We note that the results in the paper stay and become simpler if (3.2) is valid without Z_i, namely the MAR in (3.1) is valid.

In addition, we assume that the extra covariate Z does not have any predicting power on the conditional mean function of Y. This assumption is not restrictive because if a part of Z possesses such power, it should be included in X. In particular, we assume

$$E(Y|X, Z) = m(X) \quad \text{and} \quad var(Y|X, Z) = \sigma^2(X), \tag{3.3}$$

Though Z_i is invisible in the final results because of the homogeneous assumption (3.3), the missing mechanism (3.2) is important to ensure consistency of the nonparametric estimators. The key implication of (3.2) is that Y_i and δ_i are not conditional independent without given Z_i, i.e. $E(Y\delta|X) \neq m(X)w(X)$ where $w(x) = E(\delta|X = x)$.

We will consider two estimators of $m(x)$. The first estimator uses only the complete observations (those with no missing values),

$$\hat{m}_c(x) = \frac{\sum_{i=1}^n K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda})\delta_i Y_i}{\sum_{i=1}^n K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda})\delta_i}. \tag{3.4}$$

We call it the complete case estimator. As shown in the Appendix, despite the impact of a selection bias due to the missingness as specified by $w(x)$, the estimator is still consistent. The intuitive rationale of the consistency is that (3.4) is a ratio estimator so that the impacts due to selection bias arising in the numerator and denominator cancel each other. However, it does not fully utilize data information in X_i. For improvement, in the second estimator, we impute each missing Y_i by $\hat{m}_c(X_i)$, which leads to the proposed imputation based estimator

$$\hat{m}_I(x) = \frac{\sum_{i=1}^n K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda})\{\delta_i Y_i + (1-\delta_i)\hat{m}_c(X_i)\}}{\sum_{i=1}^n K_h\left(x^c - X_i^c\right) L(x^u, X_i^u; \vec{\lambda})}. \tag{3.5}$$

In a case that $m(x)$ is constant over the sample space of the covariate X, one can easily show that $\hat{m}_c(x)$ and $\hat{m}_I(x)$ are equivalent. In the following discussion, we shall focus on non-degenerate conditional mean function, i.e. the covariate X is relevant in predicting the conditional mean function. We note here that as the target of analysis is a conditional mean function, there is no need to carry out a nonparametric multiple imputation as those in Gonzalez-Manteiga and Perez-Gonzalez (2004) and Wang and Chen (2009), where the missing values are imputed for multiple times from a nonparametric estimator of the conditional distribution function. The effect of such a multiple imputation for nonparametric regression estimation with only continuous covariates is evaluated in Gonzalez-Manteiga and Perez-Gonzalez (2004), and they find that doing so does not lead to any improvement in efficiency.

4 Main Results

4.1 Effects of Imputation and Discrete Smoothing

Let $\tilde{f}(x, z)$ be the probability density function of (X_i, Z_i), $f(x) = \int f(x, z)dz$ be the marginal density of X_i, $\tilde{w}(x) = \int w(x, z)\tilde{f}(x, z)dz$ be the marginally weighted average propensity. We note that $\tilde{w}(x) \neq w(x)$ in general, which reflects the MAR assumption (3.2). We define the following with respect to the continuous kernel $K(\cdot)$: $R(K) = \int K^2(u)du$ and $\sigma_K^2 = \int u^2 K(u)du$. The following quantities are needed to describe effects of smoothing the discrete covariates. Let $\mathcal{C}_{x^u} = \{s^u : \sum_{j=1}^{d_u} I(x_j^u \neq s_j^u) = 1\}$ be the nearest strata whose discrete component differs from x^u only in one component. For a $s^u \in \mathcal{C}_{x^u}$, define

$$\alpha(x^u, s^u) = \sum_{k=1}^{d_u} c_k I(x^u_k = s^u_k) \quad \text{and} \quad \beta_\lambda(x^u, s^u) = \sum_{k=1}^{d_u} \lambda_k I(x^u_k = s^u_k)$$

where c_k is total number of the levels of X^u_{ik}. For $\hat{m}_c(x)$, the bias induced by the smoothing of the discrete variables is quantified by

$$b_{1,u}(x; \vec{\lambda}) = \sum_{s^u \in \mathcal{C}^u_x} \left(\frac{1-\beta_\lambda(x^u, s^u)}{\alpha(x^u, s^u) - 1} \left[\frac{\tilde{w}(x^c, s^u)}{\tilde{w}(x)} \{m(x^c, s^u) - m(x)\} \right] \right),$$

and the bias by smoothing the continuous variables is

$$b_{1,c}(x; h) = \tfrac{1}{2} h^2 \sigma^2_K \tilde{w}^{-1}(x) \left(\mathrm{tr}[\nabla^2 \{m(x)\tilde{w}(x)\}] - m(x)\mathrm{tr}[\nabla^2\{\tilde{w}(x)\}] \right),$$

where tr and ∇ are the trace and differentiation (with respect to x^c) operators. Similarly, the biases of $\hat{m}_I(x)$ from smoothing discrete and continuous variables are given by

$$\begin{aligned} b_{2,u}(x; \vec{\lambda}) &= \frac{f(x) - \tilde{w}(x)}{f(x)} b_{1,u}(x; \vec{\lambda}) + \sum_{s^u \in \mathcal{C}^u_x} \left(\frac{1-\beta_\lambda(x^u, s^u)}{\alpha(x^u, s^u) - 1} \left[\frac{f(x^c, s^u)}{f(x)} \{m(x^c, s^u) - m(x)\} \right] \right) \\ \text{and } b_{2,c}(x; h) &= \frac{f(x) - \tilde{w}(x)}{f(x)} b_{1,c}(x; h) + \tfrac{1}{2} h^2 \sigma^2_K f^{-1}(x) \left(\mathrm{tr}[\nabla^2\{m(x)f(x)\}] - m(x)\mathrm{tr}[\nabla^2\{f(x)\}] \right). \end{aligned}$$

Furthermore, let $V^c_1(x) = R(K)\sigma^2(x)/\tilde{w}(x)$,

$$\begin{aligned} V^I_1(x) &= \frac{\sigma^2(x)}{f^2(x)} \left[R(K)\tilde{w}(x) + 2R_2(K)\{f(x) - \tilde{w}(x)\} + R_3(K) \frac{\{f(x) - \tilde{w}(x)\}^2}{\tilde{w}(x)} \right] \text{and} \\ V^I_2(x) &= \frac{\sigma^2(x)}{f^2(x)} \left[R(K)\tilde{w}(x) + 3R_2(K)\{f(x) - \tilde{w}(x)\} + 2R_3(K) \frac{\{f(x) - \tilde{w}(x)\}^2}{\tilde{w}(x)} \right] \end{aligned}$$

where $R_2(K) = \int K^{(2)}(u)K(u)du$, $R_3(K) = \int K^{(3)}(u)K(u)du$ and $K^{(j)}(t)$ is the j^{th} convolution of $K(t)$. The following theorem, whose proof is deferred to the appendix, reports the properties of $\hat{m}_c(x)$ and $\hat{m}_I(x)$.

Theorem 1 *Under the assumptions given in the Appendix, let* $\lambda = \min_{l=1}^{d_u}\{\lambda_l\}$ *and* $A(\vec{\lambda}) = 2\sum_{j=1}^{d_u}(1 - \lambda_j)$,

$$\begin{aligned} E\{\hat{m}_c(x)\} &= m(x) + b_{1,u}(x; \vec{\lambda}) + b_{1,c}(x; h) + O\{h^2(1-\lambda)^2\}, \\ var\{\hat{m}_c(x)\} &= \frac{1}{nh^{d_c}} V^c_1(x) - \frac{A(\vec{\lambda})}{nh^{d_c}} V^c_1(x) + \frac{1}{nh^{d_c - 2}} V^c_2(x) + O[n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}], \\ E\{\hat{m}_I(x)\} &= m(x) + b_{2,u}(x; \vec{\lambda}) + b_{2,c}(x; h) + O\{h^2(1-\lambda)^2\} \quad \textit{and} \\ var\{\hat{m}_I(x)\} &= \frac{1}{nh^{d_c}} V^I_1(x) - \frac{A(\vec{\lambda})}{nh^{d_c}} V^I_2(x) + \frac{1}{nh^{d_c - 2}} V^I_3(x) + O[n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}] \end{aligned}$$

where $V^c_2(x)$ *and* $V^I_3(x)$ *are bounded terms associated with second order variances given by (A.8) and (A.15) in the Appendix.*

The implications of Theorem 1 are the following.

Remark 1. Theorem 1 shows that the imputation based estimator $\hat{m}_I(x)$ has smaller variance than $\hat{m}_c(x)$. This can be appreciated by comparing $V_1^c(x)$ and $V_1^I(x)$, which define the leading order variances of the two estimators. For commonly used symmetric kernel, 0 is the unique maximizer of K and its convolution $K^{(2)}$ and we have $R_2(K) < R(K)$ and $R_3(K) < R(K)$; see also Cheng and Peng (2006) and Cheng et al. (2007). Therefore,

$$V_1^I(x) \leq \frac{R(K)\sigma^2(x)}{f^2(x)\tilde{w}(x)}\{\tilde{w}(x) + f(x) - \tilde{w}(x)\}^2 = V_1^c(x).$$

We note here that $f(x) - \tilde{w}(x) = \int\{1 - w(x,z)\}\tilde{f}(x,z)dz \geq 0$. When there is no missing value, i.e $w(x,z) \equiv 1$, the theorem implies that the leading variance of the oracle (who knows all missing Y_i) estimator $\hat{m}(x)$ in (2.2) is

$$(nh^{d_c})^{-1}R(K)\sigma^2(x)/f(x).$$

Thus, $\hat{m}_c(x)$ which ignores the missing values endures a variance inflation by a factor $f(x)/\tilde{w}(x)$. The proposed $\hat{m}_I(x)$ removes part of the variance inflation by utilizing missing value information. Our finding here is more explicit and general than that given in Chu and Cheng (1995) who considers the case where all covariate are continuous.

As an alternative, one could replace Y_i in the numerator of (3.5) also by $\hat{m}_c(X_i)$, resulting in a smoothing twice estimator. In this case, the variance of the resulting estimator can be further reduced. However, one can show that extra bias arises due to smoothing twice. This is related to the finding in Cheng and Peng (2006) and Cheng et al. (2007), although their approaches were proposed for nonparametric regression with no missing data.

Remark 2. By smoothing the categorical variables , both $\hat{m}_c(x)$ and $\hat{m}_I(x)$ enjoy variance reductions as shown by the terms involving $A(\vec{\lambda})$. This is a result of combining data within neighboring cells defined by the discrete variables. Although the variance reductions are at the second order $(1-\lambda)/(nh^{d_c})$, the realized reduction in finite samples can be substantial, especially when large number of categorical variables in presence resulting in some cells with sparse observations which is the case for many applications. The second order result on the effects of smoothing the discrete variables can be viewed as extensions of Hall (1981) for density estimation for purely discretely-valued random variables for nonparametric regression with mixed covariates in the absence of missing values. If the regression is absent of the discrete covariates, the implication of Theorem 1 agrees to that in Chu and Cheng (1995). And we note that a second order variance reduction is the distinctive benefit of smoothing the discrete covariate. In finite sample performance, such reduction may be substantial.

Remark 3. The theorem also contains results on the bias of the estimators. The bias terms have contributions from both discrete and continuous components. By a closer look at the bias terms $b_{j,u}(\cdot)$ and $b_{j,c}(\cdot)$ where $j = 1, 2$, we may find that the feature of the bias has two sources. One is the behaviors of $m(\cdot, s^u)$, $w(\cdot, s^u)$ and $f(\cdot, s^u)$ in neighboring cells s^u around the target cell x^u; and the other is the derivatives of $m(\cdot, x^u)$, $w(\cdot, x^u)$ and $f(\cdot, x^u)$ in the local cell defined by x^u. We note that when discrete covariate is incorporated, the bias in $\hat{m}_I(\cdot)$ may not necessarily be worsen. The optimal bandwidths $(h, 1-\vec{\lambda})$ that minimize the mean square error (MSE) or the mean integrated square error (MISE) satisfy $h \sim n^{-1/(4+d_c)}$ and $(1-\vec{\lambda}) \sim n^{-2/(4+d_c)}$, which means that $\vec{A}(\lambda) \sim n^{-2/(4+d_c)}$ as well. These rates coincide with those obtained in Hall, Racine, and Li (2004) when there is no missing values.

Remark 4. When d_u is fixed and finite, we see from Theorem 1 that the variances of $\hat{m}_I(\cdot)$ and $\hat{m}_c(\cdot)$ are $O(n^{-1}h^{-d_c})$ which is coincident with that of a d_c dimensional continuous covariate smoothing. This illustrates another advantage of smoothing discrete covariate being less restrictive to the curse of dimensionality (Li and Racine, 2007). When there is no continuous covariate, the regression function estimates the overall mean of each category formed by the combinations of discrete covariate. In such case, $\hat{m}_c(x)$ and $\hat{m}_I(x)$ have the same variance of $O(n^{-1})$ in the leading order of magnitude.

4.2 Simultaneous Confidence Bands

To assess the level of uncertainty, we study how to obtain simultaneous confidence bands for the nonparametric regression estimates in the presence of discrete covariate and missing values. When all components of data are available, simultaneous confidence bands for kernel smoothing methods have been studied extensively; see for example Bickel and Rosenblatt (1973), Eubank and Speckman (1993), Xia (1998) and Zhao and Wu (2008) and reference therein.

We follow the convention in existing confidence band theory (Eubank and Speckman, 1993; Xia, 1998) by considering $d_c = 1$ and $\mathcal{X}^c = (0,1)$ for simplicity, and extend the theory to the mixed covariate case. Following Theorem 1, it can be shown following the approach in Cheng et al. (2007) that as $n \to \infty$,

$$\sqrt{\frac{nh}{V_1^I(x)}}\{\hat{m}_I(x) - m(x) - b_{2,u}(x,\vec{\lambda}) - b_{2,c}(x,h)\} \xrightarrow{d} N(0,1).$$

Thus the confidence band for $m(\cdot, x^u)$ can be constructed for each $x^u \in \mathcal{X}^u$. Because d_u is fixed and finite, for each $x^u \in \mathcal{X}^u$,

$$\lim_{n\to\infty} P\left[\sup_{x^c \in (0,1)} \left\{\sqrt{\frac{nh}{V_1^I(x)}}|\hat{m}_I(x) - m(x) - b_{2,u}(x,\vec{\lambda}) - b_{2,c}(x,h)|\right\} \le L_n(\alpha)\right] = 1-\alpha \tag{4.1}$$

where $L_n(\alpha) = \sqrt{-2\log(h)} + (A - z_\alpha)/\sqrt{-2\log(h)}$, $z_\alpha = \log\log\{(1-\alpha)^{-1/2}\}$ and $A = \log\left[\left\{\int K'^2(u)du / \int K^2(u)du\right\}^{1/2}/(2\pi)\right]$. When $\vec{\lambda} = 1$, i.e. without smoothing discrete covariate, the proof of (4.1) follows the conventional approaches in Bickel and Rosenblatt

(1973), Xia (1998) and Zhao and Wu (2008). When $\vec{\lambda}$ is chosen by the cross-validation method, we observe that the impact on the variance only occurs at the second order, so that the conventional arguments carries over for (4.1) as well. A similar result for $\hat{m}_c(x)$ is also valid. In applications, $V_1^I(x)$, $b_{2,u}(x, \vec{\lambda})$ and $b_{2,c}(x, h)$ are substituted by their consistent estimators respectively.

4.3 Locally Adaptive Smoothing

In practice, the smoothing parameter $(h, \vec{\lambda})$ can be chosen by minimizing the cross-validation score function $CV(h, \lambda) = \sum_{i=1}^n \{Y_i - \hat{m}_{h,\vec{\lambda}}^{(-i)}(X_i)\}^2 \delta_i$ where $m_{h,\vec{\lambda}}^{(-i)}(\cdot)$ is the nonparametric estimator excluding the ith using observation using $(h, \vec{\lambda})$. The h and $\vec{\lambda}$ chosen asymptotically minimize the mean integrated squared error (MISE) (Hall et al., 2004). For smoothing continuous covariate, the choice of bandwidth locally adaptive to the smoothness of the underlying regression function was studied in Fan and Gijbels (1995) and Fan, Hall, Martin, and Patil (1996).

In our study, we use the method proposed in Fan et al. (1996) to choose the bandwidth that is locally adaptive with respect to the continuous covariate. According to Theorem 1, the optimal bandwidth h is $O(n^{-1/5})$ when $d_c = 1$. Suppose there is a collection of functions G and for each $g(\cdot) \in G$, the corresponding bandwidth is given by $h(x) = n^{-1/5} g(x)$. Then the locally adaptive bandwidth can be chosen by

$$\min_{g \in G} CV(g) = \sum_{i=1}^n \left\{ Y_i - \hat{m}_{h,\vec{\lambda}}^{(-i)}(X_i) \right\}^2 \delta_i. \tag{4.2}$$

The G is chosen to be a class of continuous function with certain degrees of derivatives. According to the theoretical results in Fan et al. (1996), the minimization of the CV function over a large enough G achieves the optimal bandwidth that is adaptive to the smoothness of $m(x)$. In our paper, we follow the suggestion in Fan et al. (1996) to formulate G by cubic spline interpolation in simulation study and data analysis.

5 Simulation Studies

We conduct simulation studies to examine the finite sample performance of the methods. A natural extension of existing nonparametric regression such as in Li and Racine (2007) to situation with missing data is corresponding to the $\hat{m}_c(x)$ in our paper. Therefore in the simulation we compare the performance of $\hat{m}_c(x)$ and $m_I(x)$.

We consider three variables in the covariate with X_1 continuous following a uniform distribution on the interval $(0, 1)$, $X_2, X_3 \in \{0, 1\}$ with $P(X_2 = 0) = P(X_3 = 0) = 0.4$. The conditional mean function was chosen to be

$$E(Y|X) = m(X) = \beta_{0l} + \beta_{1l} X_1 + \beta_{2l} \sin^2\{2\pi(X_1 - 0.5)\}$$

where $l = 2X_3 + X_2 + 1$ is a one to one mapping from $\{0,1\}^2$ to $1-4$. The vector $\beta_l = (\beta_{0l}, \beta_{1l}, \beta_{2l})^T$ reflects different features with respect to the continuous variable X_1 among combinations of discrete variables. The β_l was generated from $N(\mu_1, \Sigma_1)$ and fixed throughout the simulation, where $\mu = (1, 0.5, 3.5)^T$ and $\Sigma = \text{diag}(0.04, 0.04 \times 0.5, 0.04 \times 3.5)$. The Y is generated by $Y = m(X) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. We attempted two noise levels $\sigma = 0.3$ and $\sigma = 0.6$ in the simulation study and the sample sizes were set to be $n = 50, 100, 200$. The missing propensity (no extra Z) was specified by

$$P(\delta = 1|X) = \{\exp\{-b(X)\} + 1\}^{-1} \text{ where } b(x) = \theta_{0l} + \theta_{1l}x + \theta_{2l}\phi\left(\frac{x - \theta_{3l}}{\theta_{4l}}\right)$$

and $\phi(x)$ is the density function of the standard normal distribution. The $\theta_l = (\theta_{0l}, \ldots, \theta_{4l})$ was generated from $N(\mu_2, \Sigma_2)$ and kept fixed during the simulation, and $\mu_2 = (1, 0.5, -0.3, 0.5, 0.1)^T$ and $\Sigma_2 = \text{diag}(0.01, \ldots, 0.01)$. Using this propensity function, about $30 - 35\%$ response variables are missing.

The bandwidths for smoothing discrete covariate were chosen by the Cross Validation method as in Hall et al. (2004). The bandwidth in the simulation with respect to the continuous covariate is calculated by (4.2) which is adaptive to the smoothness of the continuous covariate. The cubic spline interpolation was conducted by chosen ten equal spaced grid points. Then (4.2) is optimized with respect to the ten bandwidths. On each combination of the discrete covariate, we estimated the conditional mean function on 50 equally spaced grid points. By repeating the simulation 1000 times, we summarized the bias, variance and mean squared error (MSE) for each point. In Table 1, we report the averaged squared bias (ABias^2), variance (AVar) and MSE (AMSE) for $\hat{m}_c$ and $\hat{m}_I$ for each noise level and sample size. From Table 1, we observe that $\hat{m}_I(x)$ consistently had smaller AMSE than $\hat{m}_c(x)$ because of its smaller variance, especially when sample size is small. This confirmed our finding in Theorem 1.

To obtain the simultaneous confidence band by (4.1), we follow the approaches in Xia (1998) and Zhao and Wu (2008) to estimate $\sigma^2(x)$, $\tilde{w}(x)$ and $f(x)$ using kernel smoothing method with bandwidths chosen by minimizing corresponding cross-validation methods. The $m''(x)$ and $f'(x)$ were also estimated by kernel smoothing method whose bandwidths were chosen by the reference rule (Härdle, 1990). Two confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ were studied in the simulation whose results are reported in Table 1. It is seen from the simulation that the confidence bands had empirical coverage close to the nominal level when the sample size is reasonably large. When sample size was small, $n = 50$, the confidence bands had coverage below the nominal level. This is reasonable considering the amount of data missing and the fact of slow convergent rate of simultaneous confidence bands (Zhao and Wu, 2008).

					Coverage	
		ABias^2	AVar	AMSE	$\alpha = 0.1$	$\alpha = 0.05$
				$n = 50$		
$\hat{m}_c$	$\sigma = 0.3$	0.048	0.500	0.546	0.863	0.887
$\hat{m}_I$		0.051	0.443	0.494	0.871	0.901
$\hat{m}_c$	$\sigma = 0.6$	0.042	0.521	0.564	0.840	0.878
$\hat{m}_I$		0.050	0.465	0.515	0.848	0.899
				$n = 100$		
$\hat{m}_c$	$\sigma = 0.3$	0.017	0.135	0.152	0.904	0.953
$\hat{m}_I$		0.019	0.114	0.133	0.892	0.951
$\hat{m}_c$	$\sigma = 0.6$	0.023	0.175	0.198	0.877	0.926
$\hat{m}_I$		0.028	0.151	0.179	0.889	0.943
				$n = 200$		
$\hat{m}_c$	$\sigma = 0.3$	0.009	0.045	0.054	0.918	0.956
$\hat{m}_I$		0.011	0.038	0.049	0.916	0.959
$\hat{m}_c$	$\sigma = 0.6$	0.017	0.083	0.010	0.909	0.945
$\hat{m}_I$		0.021	0.070	0.091	0.903	0.941

Table 1: Simulation results of the averaged squared bias (ABias^2), Variance (AVar) and MSE (AMSE) of $\hat{m}_c(x)$ and $\hat{m}_I(x)$, and the empirical coverages of the confidence bands for nominal levels $\alpha = 0.1$ and $\alpha = 0.05$.

6 Data Analysis

6.1 Aids Clinical Trials Group Study 175 Data

We demonstrate the application of proposed approach by an application in the ACGT 175 data considered in Davidian, Tsiatis, and Leon (2005). The data consist of the CD4 counts along with other variables for subjects receiving ZDV (control) and three other therapies (treatments). We consider the logarithm of the CD4 counts at the 96 week as the response variable, where 1342 out of 2139 were complete cases and the rest were missing, because of death, dropout or other reasons. The logarithm of interim CD4 counts at 20 weeks is used as the continuous covariate, and the other covariates are the indicators of symptoms, prior antiretroviral therapy and Karnofsky score. Although Karnofsky score is a continuous measurement, it takes only four values (70,80,90,100) where the first two were rarely observed. So we treat it as a binary variable by dichotomizing it at 100. It is documented that the covariate can impact the missing mechanism of the response variable (Davidian et al., 2005).

Biweight kernel $K(x) = 15/16(1-x^2)^2 I(|x| \leq 1)$ is applied for continuous covariate, and we chose the smoothing bandwidths by the cross-validation, which led to $\hat{m}_c(x)$ and $\hat{m}_I(x)$. The adaptive bandwidth selected for the continuous variable is close to a constant, due to the lack of local feature as seen from the estimated regression function. We then estimated the density function $f(x)$, the marginal propensity function $\tilde{w}(x)$, the conditional variance function $\sigma^2(x)$ and the relevant derivatives by the kernel smoothing methods respectively as in the simulation study. Based on these estimated quantities, we obtained the confidence bands from (4.1). Figure 1 displays the nonparametric regression estimates $\hat{m}_c(x)$ and $\hat{m}_I(x)$ based on the complete cases and the imputation respectively and the associated 95% confidence bandiwdths. We observe from Figure 1 that heterogeneous patterns of the conditional mean function with respect to the interim CD4 counts. The estimated regression functions differ between the control and treatment groups, as well as among different combinations of the other discrete covariate. The feature of the regression function found by the nonparametric method is complementary to the additive parametric model with quadratic predictors considered in Davidian et al. (2005). The $\hat{m}_c(x)$ and $\hat{m}_I(x)$ estimates share some similar pattern, though there are some differences at various places. We note that the width of the confidence bands for $\hat{m}_I(x)$ is smaller than that of $\hat{m}_c(x)$, which is expected from our theoretical results. From the confidence band estimates, we observed a widening trend at the two ends for the two estimates. This is mainly due to the sparsity of observations towards the both ends.

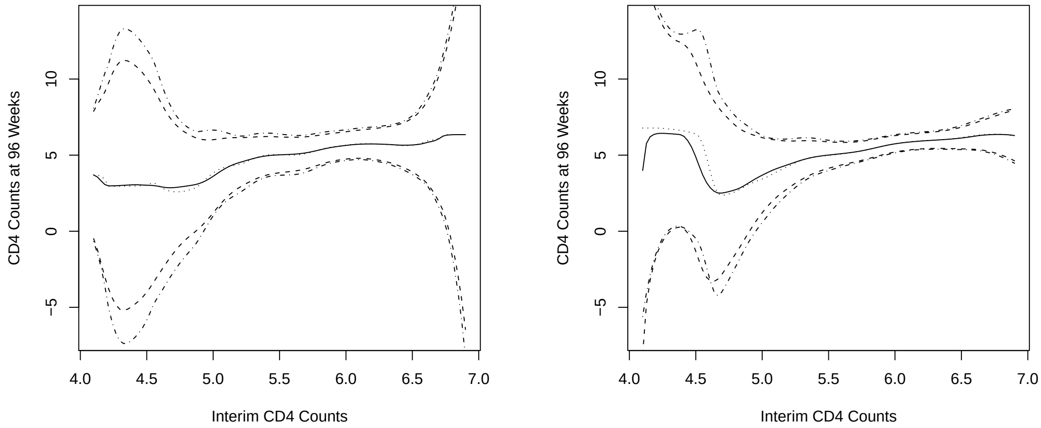

Figure 1: The $\hat{m}_c(x)$ (dotted lines) and $\hat{m}_I(x)$ (solid lines) for the controlled (left panel) and treated (right panel) subjects with HIV symptoms, experienced prior antiretroviral treatment and Karnofsky score 100. The dash dotted lines and the dash lines are the estimated 95% confidence bands for $\hat{m}_c(x)$ and $\hat{m}_I(x)$.

6.2 Estimating US Census Enumeration Probability Function

We apply the proposed kernel estimators to an ACE revision II research data (US Census Bureau, 2004) of the 2000 US Census. The ACE study consists of two independent samples from the US population, namely the E and P samples respectively. The capture-recapture design is implemented to identify matches between the two samples. The response variable Y is the match indicator and its conditional mean given the covariate is the important enumeration probability function. A significant portion of enumerations in the US Census ACE study has missing match indicator, due to the fact that a definite match or no-match can not be established between the records in the P and E samples. In the ACE revision II research data that we will analyze in this paper, missing response accounts for 3.7% of the enumerations in the P sample and 6.5% in the E sample. These are high comparing with the overall level of undercounts in the US Census (US Census Bureau, 2004). The data contain about $60,000$ P-sample cases and $70,000$ E-sample cases, which makes an ideal case for the nonparametric regression estimation.

In our analysis, the covariates X_i include age (X_{i1}), sex (X_{i2}, 2 levels), housing tenure (X_{i3}, 2 levels: owner and renter), and racial origins (X_{i4}, 7 levels: American Indian or Alaska Natives on Reservation, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian and Non-Hispanic white or other races) and Geographical region (X_{i5}, 4 levels: Northeast, Midwest, South and West). Additional covariates may be included without changing the tune of the analysis. The response variable Y_i is the match status between the P sample $\mathcal{P}$ and the E sample $\mathcal{E}$, where

$$Y_i = \begin{cases} 1\text{(a match)} & \text{if } i \in \mathcal{P} \cap \mathcal{E} \ ; \\ 0\text{(not a match)} & \text{if } i \in \mathcal{P} \text{ but not in } \mathcal{E}; \\ ?\text{(missing value)} & \text{if } i \text{ is an unresolved case.} \end{cases} \tag{6.1}$$

The interest is in estimating $m(x) = E(Y_i|X_i = x)$, the enumeration probability function.

We chose the biweight kernel $K(x) = 15/16(1 - x^2)^2 I(|x| \leq 1)$ to smooth the age and the discrete kernel (2.1) to smooth the other categorical covariates. The smoothing bandwidths were chosen by the cross-Validation (CV) method. In this simplified analysis and without loss of general focus, we let all the discrete bandwidths λ_i being equal to one λ. In a comprehensive case study of the Census ACE data (Chen, Tang and Mule, 2010), a dedicated two stage bandwidth selection procedure is implemented. Let $\hat{m}_{h,\lambda}^{(-i)}(x)$ be the estimators of $m(x)$ after excluding the i^{th} data pair (X_i, Y_i). Specifically, we define the cross-validation score based on all complete records

$$CV(h, \lambda) = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{m}_{h,\lambda}^{(-i)}(X_i)\}^2 \delta_i. \tag{6.2}$$

The bandwidths prescribed were $h = 5.5$ and $\lambda = 0.8$, which were the bandwidths used in the imputation based estimates for $m(x)$ in Figure 2.

It is observed from Figure 2 that the covariate in the analysis contribute to the heterogeneity in the enumeration probability function $m(x)$. The age effect was quite apparent in the estimates for $m(x)$. At the same time, the kernel estimates changes substantially with respect to the other categorical variables. Figure 2 indicates that Northeast White Male Owner had an overall higher enumeration probability than Northeast Hispanic Female owners and Midwest Black Male renters, which might be expected. While these confirm the effects of these covariates, they do reveal the difficulty in capture the underlying forms of the functions with respect to these discrete covariates. The wave-like pattern in the $m(x)$ estimates in some cells suggests some age-heaping in a multiple of 5 or 10 years in age beyond 30. Figure 3 displays the kernel estimates for the missing propensity score $w(x)$, which was as interesting as Figure 2. For instance, the White Male owners had very small chance of being missing. In contrast, the Hispanic Female owners endured larger missingness while the Black Male renters experienced the highest missing values among the three.

Both Figures 2 and 3 also reveal challenges that one would face in proposing a reasonable parametric regression models. There are 112 post-strata based on the four discrete covariates. The sample size within some of these 112 post-strata can be very small, for instance the Native Hawaiian or Pacific slander. Getting a workable model for each stratum is quite a task. The task will only grow when more covariates are included. At the same time, the figures show that the proposed the kernel estimation is flexible and adaptive to varying functional forms in $m(x)$. As shown by our theoretical investigation, the kernel estimates are consistent and reflective to the underlying model structure without imposing any subjective assumptions.

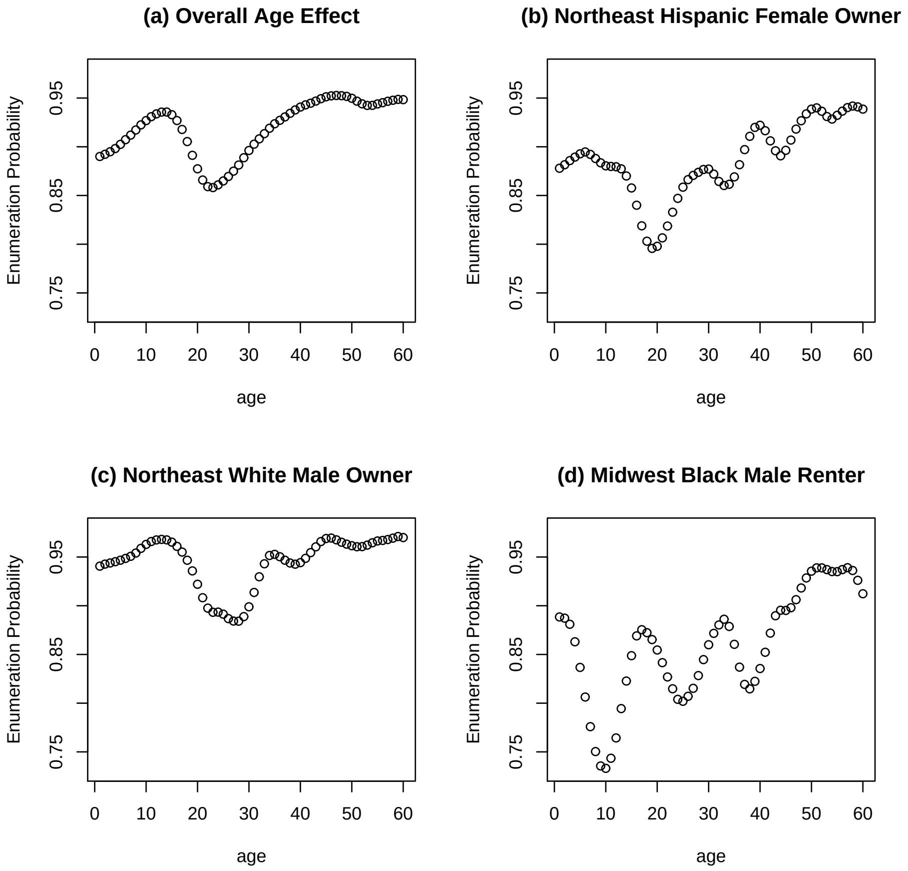

Figure 2: Kernel estimates of the enumeration probability $m(x)$ based on $\hat{m}_I(x)$. Bandwidths used are $h = 5.5$ and $\lambda = 0.8$.

Appendix: Technical Details

Technical Assumptions

Let covariate $X_i = (X_i^c, X_i^u)$ where X_i^c is a d_c-dimensional continuous covariate and X_i^u is a d_u-dimensional unordered categorical covariate, $\mathcal{X} = \{\mathcal{X}^c, \mathcal{X}^u\}$ be the support of X_i, where $\mathcal{X}^c$ and $\mathcal{X}^u$ are the supports of X_i^c and X_i^u respectively. We assume the model (3.3) for independent and identically distributed data pairs $\{(X_i, Z_i, Y_i)\}_{i=1}^n$. And the following conditions are assumed in Theorem 1.

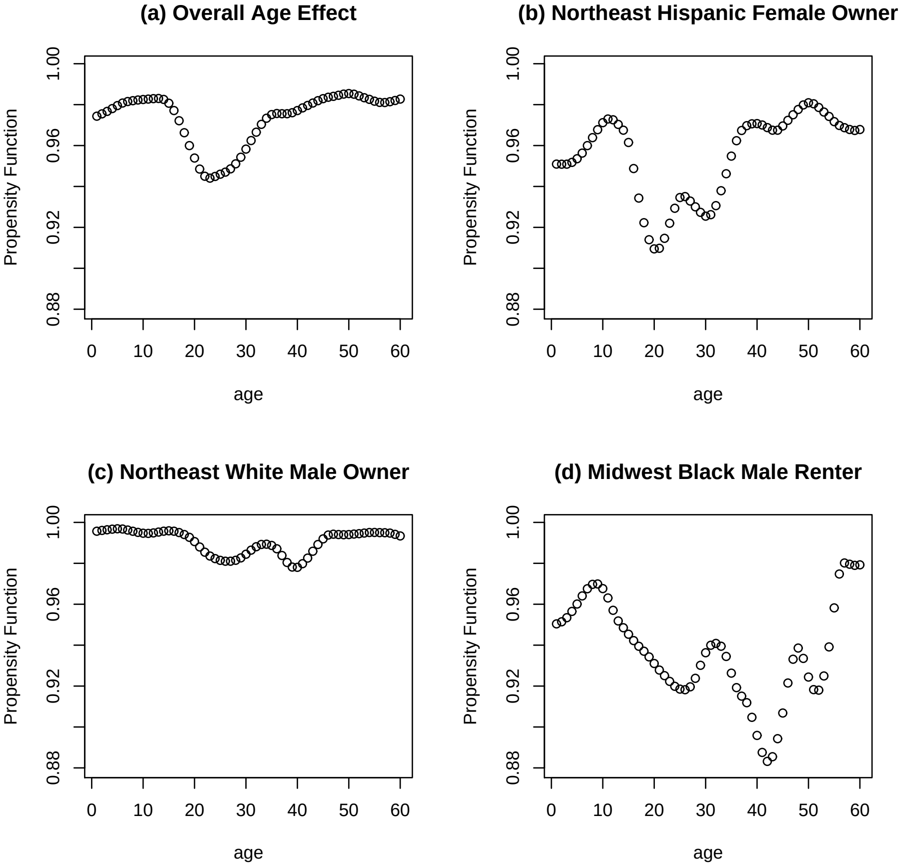

Figure 3: Kernel estimates of the census missing propensity function $w(x)$. Bandwidths used are $h = 5.0$ and $\lambda = 0.8$.

C.1 Let $K(\cdot)$ be a d_c variates nonnegative, bounded and symmetric probability density function with bounded second derivative. The smoothing bandwidths satisfy that $h \to 0$ and $\max\limits_{1\le j\le d_u}\{(1-\lambda_j)\} \to 0$ and $nh^{d_c}/\log(n) \to \infty$ as $n \to \infty$.

C.2 We assume missing at random in Y, namely $P(\delta = 1|Y, X, Z) = P(\delta = 1|X, Z) := w(X, Z)$, where $w(x^c, x^u, z) \ge C_w > 0$ for a constant C_w and $w(x^c, x^u, z)$ has bounded continuous second partial derivative with respect to x^c.

C.3 For given $x^u \in \mathcal{X}^u$, $m(x^c, x^u)$ and the probability density of the covariate $f(x^c, x^u)$ have bounded continuous second partial derivatives with respect to x^c, and there exist $C_f > 0$ such that $f(x^c, x^u) \ge C_f$.

Sketched Proof of Theorem 1.

We sketch the proof of Theorem 1 here where the complete proof with more details is available in Tang (2008). Let $\mathcal{K}_{h,\vec{\lambda}}(u, v) = K_h(u^c - v^c)\, L(u^c, v^u; \vec{\lambda})$ and define $\hat{f}_c(x) =$

$n^{-1}\sum_{i=1}^{n}\mathcal{K}_{h,\vec{\lambda}}(x,X_i)\delta_i$ and $\hat{\phi}_c(x)=n^{-1}\sum_{i=1}^{n}\mathcal{K}_{h,\vec{\lambda}}(x,X_i)\delta_i Y_i$. From (3.4), $\hat{m}_c(x)=\hat{\phi}_c(x)/\hat{f}_c(x)$. We show that

$$
\begin{aligned}
E\{\hat{f}_c(x)\} &= \tilde{w}(x)+\tfrac{1}{2}h^2\sigma_K^2\mathrm{tr}[\nabla^2\{\tilde{w}(x)f(x)\}]-\tilde{w}(x)\textstyle\sum_{j=1}^{p_u}(1-\lambda_j)\\
&\quad+\sum_{s^u\in\mathcal{C}_x^u}\frac{1-\beta_\lambda(x^u,s^u)}{\alpha(x^u,s^u)-1}\tilde{w}(x^c,s^u)+O\{h^2(1-\lambda)^2\}, && \text{(A.1)}\\
var\{\hat{f}_c(x)\} &= \frac{1-A(\vec{\lambda})}{nh^{d_c}}R(K)\tilde{w}(x)+\frac{1}{2nh^{d_c-2}}S_2(K)\mathrm{tr}[\nabla^2\{\tilde{w}(x)f(x)\}]\\
&\quad+O(n^{-1}h^{-d_c+2}\{h^2+(1-\lambda)^2\}), && \text{(A.2)}\\
E\{\hat{\phi}_c(x)\} &= m(x)\tilde{w}(x)+\tfrac{1}{2}h^2\sigma_K^2\mathrm{tr}[\nabla^2\{m(x)\tilde{w}(x)\}]-m(x)\tilde{w}(x)\textstyle\sum_{j=1}^{p_u}(1-\lambda_j)\\
&\quad+\sum_{s^u\in\mathcal{C}_x^u}\frac{1-\beta_\lambda(x^u,s^u)}{\alpha(x^u,s^u)-1}m(x^c,s^u)\tilde{w}(x^c,s^u)+O\{h^2(1-\lambda)^2\}\quad\text{and} && \text{(A.3)}\\
var\{\hat{\phi}_c(x)\} &= \frac{1-A(\vec{\lambda})}{nh^{d_c}}R(K)\tilde{w}(x)\{m^2(x)+\sigma^2(x)\}+\frac{S_2(K)\mathrm{tr}\left(\nabla^2[\{m^2(x)+\sigma^2(x)\}\tilde{w}(x)f(x)]\right)}{2nh^{d_c-2}}\\
&\quad+O(n^{-1}h^{-d_c+2}\{h^2+(1-\lambda)^2\}), && \text{(A.4)}
\end{aligned}
$$

where $S_i(K)=\int u^iK^2(u)du$. Results in (A.1)-(A.4) indicate that based on the complete data only, the estimation of the density function $f(x)$ by $\hat{f}_c(x)$ is actually biased and so is the estimation of $\phi(x)=m(x)f(x)$ by $\hat{\phi}_c(x)$. Similarly, we have

$$
\begin{aligned}
cov\{\hat{\phi}_c(x),\hat{f}_c(x)\} &= \frac{1-A(\vec{\lambda})}{nh^{d_c}}R(K)m(x)\tilde{w}(x)+\frac{S_2(K)\mathrm{tr}[\nabla^2\{m(x)\tilde{w}(x)f(x)\}]}{2nh^{d_c-2}}\\
&\quad+O(n^{-1}h^{-d_c+2}\{h^2+(1-\lambda)^2\}). && \text{(A.5)}
\end{aligned}
$$

Define $f_c(x)=\tilde{w}(x)$ and $\phi_c(x)=m(x)\tilde{w}(x)$. An expansion of $\hat{m}_c(x)$ is

$$
\hat{m}_c(x)=\frac{\phi_c(x)}{f_c(x)}+\frac{\{\hat{\phi}_c(x)-\phi_c(x)\}}{f_c(x)}-\frac{\phi_c(x)\{\hat{f}_c(x)-f_c(x)\}}{f_c^2(x)}\{1+o_p(1)\}. \qquad \text{(A.6)}
$$

Taking expectation on (A.6), and from (A.1) and (A.3),

$$
E\{\hat{m}_c(x)\}=m(x)+b_{1,u}(x;\vec{\lambda})+b_{1,c}(x;h)+O\{h^2(1-\lambda)^2\}. \qquad \text{(A.7)}
$$

By applying the variance operation on (A.6) and substituting equations (A.2), (A.4) and (A.5), we summarize that

$$
\begin{aligned}
var\{\hat{m}_c(x)\} &= \frac{1}{nh^{d_c}}V_1^c(x)-\frac{A(\vec{\lambda})}{nh^{d_c}}V_1^c(x)+\frac{1}{nh^{d_c-2}}V_2^c(x)+O(n^{-1}h^{-d_c+2}\{h^2+(1-\lambda)^2\})\\
\text{and } V_2^c(x) &= \frac{1}{2\tilde{w}^2(x)}\big\{S_2(K)\big(\mathrm{tr}\,\big[m^2(x)\nabla^2\{\tilde{w}(x)f(x)\}+\nabla^2[\{m^2(x)+\sigma^2(x)\}\tilde{w}(x)f(x)]\\
&\quad-\ 2m(x)\nabla^2\{m(x)\tilde{w}(x)f(x)\}\big]\big)\big\}. && \text{(A.8)}
\end{aligned}
$$

These establishes the first part of Theorem 1.

We now evaluate the imputation based estimator $\hat{m}_I(x)$. Let $\hat{f}(x) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_{h,\vec{\lambda}}(x, X_i)$ and $\hat{\phi}_i(x) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_{h,\vec{\lambda}}(x, X_i)(1-\delta_i)\hat{m}_c(X_i)$. From (3.5), by letting $\hat{\phi}_I(x) = \hat{\phi}_c(x) + \hat{\phi}_i(x)$,

$$\hat{m}_I(x) = \{\hat{\phi}_c(x) + \hat{\phi}_i(x)\}/\hat{f}(x) = \hat{\phi}_I(x)/\hat{f}(x). \tag{A.9}$$

Similar to (A.6), we may establish an expansion for (A.9). Let $\phi_I(x) = m(x)f(x)$, we have

$$\hat{m}_I(x) = \frac{\phi_I(x)}{f(x)} + \frac{\{\hat{\phi}_I(x) - \phi_I(x)\}}{f(x)} - \frac{\phi_I(x)\{\hat{f}(x) - f(x)\}}{f^2(x)}\{1 + o_p(1)\}. \tag{A.10}$$

We establish that

$$\begin{aligned} E\{\hat{f}(x)\} &= f(x) + \tfrac{1}{2}h^2\sigma_K^2 \mathrm{tr}[\nabla^2\{f(x)\}] - f(x)\textstyle\sum_{j=1}^{p_u}(1-\lambda_j) \\ &\quad + \sum_{s^u \in \mathcal{C}_x^u} \frac{1-\beta_\lambda(x^u, s^u)}{\alpha(x^u, s^u) - 1} f(x^c, s^u) + O\{h^2(1-\lambda)^2\} \quad \text{and} \\ var\{\hat{f}(x)\} &= \frac{1 - A(\vec{\lambda})}{nh^{d_c}} R(K) f(x) + \frac{1}{2nh^{d_c-2}} S_2(K) \mathrm{tr}[\nabla^2\{f(x)\}] \\ &\quad + O(n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}). \end{aligned} \tag{A.11}$$

By taking expectation on $\hat{\phi}_I(x)$, we have

$$E\{\hat{m}_I(x)\} = m(x) + b_{2,u}(x; \vec{\lambda}) + b_{2,c}(x; h) + O\{h^2(1-\lambda)^2\}.$$

The following notations of functions are introduced for simplification. Let $g_1 = \frac{1-\tilde{w}}{f\tilde{w}^2}$, $g_2 = m\tilde{w}f$, $g_3 = \frac{1-\tilde{w}}{\tilde{w}}$, $g_4 = \{m^2 + \sigma^2\}\tilde{w}f$, $g_5 = \tilde{w}f$, $g_6(x) = \frac{\{1-\tilde{w}\}m}{\tilde{w}}$ and $g_7(x) = \frac{\{1-\tilde{w}\}m}{f\tilde{w}^2}$. Further, it can be shown that

$$\begin{aligned} var\{\hat{\phi}_i(x)\} &= \frac{1}{nh^{d_c}} \left[R(K)\{f(x) - \tilde{w}(x)\}m^2(x) + R_3(K)\sigma^2(x)\frac{\{f(x) - \tilde{w}(x)\}^2}{\tilde{w}(x)} \right] \\ &\quad - \frac{A(\vec{\lambda})}{nh^{d_c}} \left[R(K)\{f(x) - \tilde{w}(x)\}m^2(x) + 2R_3(K)\sigma^2(x)\frac{\{f(x) - \tilde{w}(x)\}^2}{\tilde{w}(x)} \right] \\ &\quad + \frac{\xi_{12}(K)R(K)}{nh^{d_c-2}} \left(g_1 g_2 \mathrm{tr}[\nabla^2\{g_2\}] - g_7 g_2 \mathrm{tr}[\nabla^2\{g_5\}] \right) \\ &\quad + \frac{1}{nh^{d_c-2}} \left\{ \tfrac{1}{2} S_2(K) C_1 + \xi_{32}(K) C_2 + \zeta_2(K) C_3 \right\} \\ &\quad + O(n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}), \end{aligned} \tag{A.12}$$

where $\xi_{ij}(K) = \int u^j K^{(i)}(u)K(u)du$ and $\zeta_i(K) = \int u_1 u_2 K^{(i)}(u_1 + u_2)K(u_1)K(u_2)du_1 du_2$,
$C_1(x) = \mathrm{tr}\left(\nabla^2[\{1-\tilde{w}\}\{m^2(x)f]\right)$,
$C_2(x) = \{1-\tilde{w}\}\sigma^2(x)f\mathrm{tr}[\nabla^2\{g_3\}] + \frac{1}{2}g_3^2\mathrm{tr}[\nabla^2\{g_4\}] + \frac{1}{2}g_6^2\mathrm{tr}[\nabla^2\{g_5\}] - g_3 g_6 \mathrm{tr}[\nabla^2\{g_5\}]$,
$C_3(x) = 2g_3\nabla^T\{g_3\}J\nabla\{g_4\} + g_6\nabla^T\{g_6\}J\nabla\{g_5\} + g_4\nabla^T\{g_3\}J\nabla\{g_3\} + g_5\nabla^T\{g_6\}J\nabla\{g_6\}$ $+g_6\nabla^T\{g_5\}J\nabla\{g_5\} - 2[g_6\nabla^T\{g_2\}J[\nabla\{g_3\} + g_2\nabla^T\{g_3\}J\nabla\{g_6\} + g_3\nabla^T\{g_2\}J\nabla\{g_6\}]$. Here $J = \mathbf{1}\mathbf{1}^T$ for $\mathbf{1} = (1, \ldots, 1)^T_{1\times d_c}$. In a similar fashion, we establish that

$$\begin{aligned} cov\{\hat{\phi}_c(x), \hat{\phi}_i(x)\} &= \frac{\{1 - 1.5A(\vec{\lambda})\}}{nh^{d_c}} R_2(K)\{f(x) - \tilde{w}(x)\}\sigma^2(x) \\ &\quad + \frac{1}{nh^{d_c-2}}\{\zeta_1(K)C_4 + \tfrac{1}{2}\xi_{22}(K)C_5\} + O(n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}) \end{aligned} \tag{A.13}$$

where $C_4(x) = \nabla^T\{g_3\}J\nabla\{g_4\} - \nabla^T\{g_2\}J\nabla\{g_6\}$ and $C_5(x) = g_4\mathrm{tr}[\nabla^2\{g_3\}] + g_3\mathrm{tr}[\nabla^2\{g_4\}] - g_2\mathrm{tr}[\nabla^2\{g_6\}] - g_6\mathrm{tr}[\nabla^2\{g_2\}]$. Furthermore

$$\begin{aligned} cov\{\hat{\phi}_I(x), \hat{f}(x)\} &= \frac{1 - A(\vec{\lambda})}{nh^{d_c}} R(K)\left\{m(x)\{f(x) - \tilde{w}(x)\} + m(x)\tilde{w}(x)\right\} \\ &+ \frac{\xi_{12}(K)S_0(K)}{nh^{d_c-2}}\left(g_3(x)\mathrm{tr}[\nabla^2\{g_2\}] - g_6\mathrm{tr}[\nabla^2\{g_5\}]\right) \\ &+ \frac{1}{nh^{d_c-2}}\left\{\tfrac{1}{2}S_2(K)C_6 + \zeta_1(K)C_7 + \tfrac{1}{2}\xi_{22}(K)C_9\right\} \\ &+ O(n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}), \end{aligned} \tag{A.14}$$

where $C_6(x) = \mathrm{tr}\left(\nabla^2[\{1 - w\}mf]\right) + \mathrm{tr}\left[\nabla^2\{g_2\}\right]$, $C_7(x) = \nabla^T\{g_2\}J\nabla\{g_3\} - \nabla^T\{g_5\}J\nabla\{g_6\}$ and $C_9(x) = g_2\mathrm{tr}[\nabla^2\{g_3\}] + g_3\mathrm{tr}[\nabla^2\{g_2\}] - g_5\mathrm{tr}[\nabla^2\{g_6\}] - g_6\mathrm{tr}[\nabla^2\{g_5\}]$.

Finally, from (A.4), (A.11), (A.12), (A.13) and (A.14), we have

$$\begin{aligned} var\{\hat{m}_I(x)\} &= \frac{1}{nh^{d_c}}V_1^I(x) - \frac{A(\vec{\lambda})}{nh^{d_c}}V_2^I(x) + \frac{1}{nh^{d_c-2}}V_3^I(x) + O(n^{-1}h^{-d_c+2}\{h^2 + (1-\lambda)^2\}), \\ V_3^I(x) &= S_2(K)T_1(x) + \xi_{22}(K)T_2(x) + \xi_{32}(K)T_3(x) + \zeta_1(K)T_4(x) + \zeta_2(K)T_5(x) \end{aligned} \tag{A.15}$$

where $T_1(x) = \frac{1}{2}f^{-2}(x)\mathrm{tr}\left[\nabla^2\{g_4\} + C_1(x) + m^2(x)\nabla^2\{f(x)\} - 2m(x)C_6(x)\right]$, $T_2(x) = \{C_5(x) - m(x)C_9(x)\}/f^2(x)$, $T_3(x) = C_2(x)/f^2(x)$, $T_4(x) = 2\{C_4(x) - m(x)C_7(x)\}/f^2(x)$ and $T_5(x) = \frac{C_3(x)}{f^2(x)}$. These conclude Theorem 1.

References

Aitchison, J. and Aitken, C. (1976), "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. (1993), "Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussions)," *Journal of the American Statistical Association*, 88, 1149–1166.

Bickel, P. J. and Rosenblatt, M. (1973), "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, 1, 1071–1095.

Chen, S. X., Tang, C. Y., and Mule, V. T. (2010), "Local post-stratification and diagnostics in dual system accuracy and coverage evaluation for the U.S. Census," *Journal of the American Statistical Association*, 105, 105–119.

Cheng, M. and Peng, L. (2006), "Simple and efficient improvement of multivariate local linear regression," *Jourmal of Multivariate Analysis*, 97, 1501–1524.

Cheng, M., Peng, L., and Wu, J. (2007), "Reducing variance in univariate smoothing," *The Annals of Statistics*, 35, 522–542.

Cheng, P. E. (1994), "Nonparametric-estimation of mean functionals with data missing at random," *Journal of the American Statistical Association*, 89, 81–87.

Chu, C. K. and Cheng, P. E. (1995), "Nonparametric regression estimation with missing data," *Journal of Statistical Planning and Inference*, 48, 85–99.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005), "Semiparametric estimation of treatment effect in a pretest-posttest study with missing data," *Statistical Science*, 20, 261–301.

Efromovich, S. (2011), "Nonparametric regression with responses missing at random," *Journal of Statistical Planning and Inference*, 141, 3744–3752.

Eubank, R. L. and Speckman, P. L. (1993), "Confidence bands in nonparametric regression," *Journal of the American Statistical Association*, 88, 1287–1301.

Fan, J. and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth an spatial adaptation," *Journal of the Royal Statistical Society, Series B*, 57, 371–394.

— (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

Fan, J., Hall, P., Martin, M. A., and Patil, P. (1996), "On local smoothing of nonparametric curve estimators," *Journal of the American Statistical Association*, 91, 258–266.

Gonzalez-Manteiga, W. and Perez-Gonzalez, A. (2004), "Nonparametric mean estimation with missing values," *Communication in Statistics*, 33, 277–303.

Hall, P. (1981), "On nonparametric multivariate binary discrimination," *Biometrika*, 68, 287–294.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hall, P. and Wand, M. P. (1988), "On nonparametric discrimination using density differences," *Biometrika*, 75, 541–547.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Li, Q. and Racine, J. S. (2003), "Nonparametric estimation of distributions with both categorical and continuous data," *Journal of Multivariate Analysis*, 86, 266–292.

— (2007), *Nonparametric Econometrics, Theory and Practice*, Princeton University Press.

Little, R. and Rubin, D. (2002), *Statistical Analysis With Missing Data*, Wiley, 2nd ed.

Müller, U. U. (2009), "Estimating linear functionals in nonlinear regression with respones missing at random," *The Annals of Statistics*, 37, 2245–2277.

Qin, J., Leung, D., and Shao, J. (2002), "Estimation with survey data under nonignorable nonresponse or informative sampling," *Journal of the American Statistical Association*, 97, 193–200.

Qin, J., Zhang, B., and Leung, D. H. Y. (2009), "Empirical likelihood in missing data problem," *Journal of the American Statistical Association*, 104, 1492–1503.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D. B. (1976), "Inference and missing values (with discussion)," *Biometrika*, 63, 581–592.

Scott, D. W. (1992), *Multivarite Density Estimation: Theory, Practice, and Visualization*, Wiley.

Tang, C. Y. (2008), "Parameter estimation and bias correction for diffusion processes, and, A nonparametric approach to census population size estimation," *Ph.D Disseration (http://gradworks.umi.com/33/10/3310803.html). Iowa State University.*

Titterington, D. M. and Mill, G. M. (1983), "Kernel-based density estimates from incomplete data," *Journal of the Royal Statistical Society, Ser B*, 45, 258–266.

US Census Bureau (2004), *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*, US Census Bureau.

Wang, D. and Chen, S. X. (2009), "Empirical likelihood for estimating equations with missing values," *The Annals of Statistics*, 37, 490–517.

Xia, Y. (1998), "Bias-corrected confidence bands in nonparametric regression," *Journal of the Royal Statistical Society, Series B*, 60, 797–811.

Zhao, Z. and Wu, W. B. (2008), "Confidence bands in nonparametric time series regression," *The Annals of Statistics*, 36, 1854–1878.